Cluster Analysis of Stocks

Project Proposal Paper

Jordan Zane Caleb Clough Matt Dyer Evan Linden

ABSTRACT

The aim of the *Cluster Analysis of Stocks* was to mine extensive longitudinal American stock/etf market data. Through this process, the project seeks to answer the questions: Do greater nuance and cross relationships, outside of typical stock groupings (technology, manufacturing, finance, etc.), exist? If so, do they allow for novel (or atleast modified) classification grouping schemes? And are there possible explanations for these groupings?

Our results dictated a combination of interesting yet verifiable knowledge. The resulting clusters, grouped by their price movement, show that stocks aren't necessarily related by sector but actually more of a particular theme. Stocks move in interesting ways yet the clustering shows key similarities within the stocks. For example, we found that a certain cluster contained a great magnitude of luxury item companies. This is notably interesting because this particular cluster fluctuates with how the economy is in nature. The key aspects of our knowledge gained is that it makes sense why some of these clusters move together and also takes stabs at the fundamental groupings that stocks have.

INTRODUCTION

To be specific, the Cluster Analysis of Stocks aims to cluster individual stocks into groups to see what individual stocks tend to "move" together. This project is not concerned with classic stock groupings, but is instead concerned with using a wider perspective on the stock market in order to mine relationships between stocks that are not logical or common knowledge.

Potential applications of the results of this project could be to more accurately predict stock movements, or atleast correlations between certains stocks/etf that are found to be in similar groups. Not only could this data be useful for identifying undervalued assets, but it could also be useful to hedge risk in a portfolio by ensuring multiple cluster groups are represented in a portfolio--ensuring groups tend to move independently, while having multiple groups present in a portfolio can ensure the overall portfolio balance does not swing wildly.

RELATED WORK

Studies come from Kaggle:

- Which stocks should I invest in? - included daily percentage change in price and volume, Correlation of percent change, Predicting expected gain and loss for each ticker, and growth trends of tickers
- Stock diversity analysis - analysing clusters to identify stocks with opposite trends
- Linear regression of stock histories - using a regression model to predict stock prices
- Wsb stock analysis - using popular tickers mentioned on reddit.com to analyze growt

DATA SET

Our main dataset is a .csv containing 5 years of AMEX, NYSE, and NASDAQ end of day data. The source of this data is Kaggle and the URL is as follows: https://www.kaggle.com/qks1lver/amex-nyse-nasdaq-stock-histories?select=fh_5yrs.csv

The overall size of the dataset is 712 MB, with a total of 6852039 data points. The data is structured such that each stock ticker represented has (up to) 5 years of daily data.

Each data point has a total of 8 attributes. They are as follows, with information about the type of variable each is:

- Date - Ordinal
- Volume - Numeric (Ratio)
- Open - Numeric (Ratio)
- High - Numeric (Ratio)
- Low - Numeric (Ratio)
- Close - Numeric (Ratio)
- Adj Close - Numeric (Ratio)
- Symbol - Categorical

The data appears to be very clean, but likely some minor cleaning will need to be done, as described in the aforementioned sections.

Additionally, we have a secondary data set to supplement the main dataset in the event that the primary dataset is not sufficient to meet our project objectives. This secondary dataset can be

found at: https://www.kaggle.com/borismarjanovic/price-volume-data-for-all-us-stocks-etfs

This dataset is structured a bit differently than the primary dataset. Each stock and etf has a separate txt file with longitudinal data. Each txt file is structured with the data being comma delimited, and thus can be treated as a csv with minimal modification. In total there are 8539 files totaling 771 MB. Each file contains a total of 7 attributes. They are as follows, with information about the type of variable each is:

- Date - Ordinal
- Open - Numeric (Ratio)
- High - Numeric (Ratio)
- Low - Numeric (Ratio)
- Close - Numeric (Ratio)
- Volume - Numeric (Ratio)
- Open Int - Numeric (Ratio)

This data also appears to be very clean, but may require substantial post-processing to integrate with the primary dataset in the event that it is needed

MAIN TECHNIQUES APPLIED

After we had collected our complete data set from Kaggle, we had to clean the data before we did any analysis with it. The data cleaning process started by dropping any NULL values from the dataset. Then the next step was to replace any string values such as '$' with a ' ', so we did not have any strings in our data set. Considering we are going to be comparing the stocks based on their prices over time, we also needed to convert the dates using the Pandas function "to_datetime", so that Python recognizes them as Dates, not strings. The last step of data cleansing was to locate and remove any outliers from the data set. Once these steps were completed, the dataset had been cleansed of any imperfections and is now primed for any data-mining we pursue later on .

First we measured daily stock movement by subtracting the closing price of one day from the closing price of the previous day for each stock for every day. If the stock increased, we assigned that day as a 1 and assigned that day as a 0 otherwise. Since we were dealing with several days of thousands of stocks, the data were already highly dimensional. We truncated the data by removing most of the attributes aside from the aforementioned binary variable measuring daily movement of the stock.

The clustering was performed using k-means clustering in the sklearn.cluster library in Python. Euclidean distance was used as the similarity measure and mean distance will be the metric to determine the distance between clusters. We used the elbow method to determine the optimal number of clusters by measuring the distortion score as a function of the number of clusters. The distortion score showed the sum of squared differences of each point to the center of its cluster. We found the distortion values using kmeans.inertia and used our judgment to find the inflection point to identify the point of diminishing returns. With thousands of stocks at our disposal, we wanted to avoid having too many clusters. After all, our goal was to find interesting and less obvious groups. Therefore, we want to avoid overfitting the data.

For data integration, we used multiple stock datasets containing similar attributes. To avoid redundancy in the data, the standard stock symbols were used for entity identification. The dates were converted to datetime using the Pandas library in Python to ensure that all dates are in a standard tractable format. This ensured that no duplicate dates were used for the same stock.

Our key technique for separating useful data from older data was by using a data cube. We split the data into different arrays to separate out the data. The most important factor of our data cube was to find relevant pricing information by using more current dates, and after slicing up all the data into a cube we decided that we were going to take the data that had been in the most recent 2 years. This allowed us to get more of an accurate prediction and also made it so we still had historic data of the stock ticker to analyze it's day-to-day trends and growth even further.

When we first looked at our data after cleaning it we decided that using a clustering classification method would work best for our situation. In our case we could set up different clusters based on how the ticker performed from it's previous day in a day-to-day setting. By using this method we created some code that marked the stock as a 1 if it closed higher than the previous day and a 0 if it did not. By doing this we could easily set up a k-means cluster to cluster the stocks that perform similarly together. We figured out the number of clusters that would be optimal by using the elbow method in which we plotted a graph of Distortion as a function of K and found out that our optimal number would be 48. We then used python code to cluster the stocks and had our results.

KEY RESULTS

1. Pre-Processed Results

As previously mentioned, the main goal of our current efforts has been to preprocess our main data csv such that a clustering analysis can be performed effectively. In order to simplify the data to allow for easier training, the group decided to identify instances where a certain stock has increased day-over-day, with a simple 1 (yes) or 0 (no) scheme. This will allow us to then do a market-basket type analysis and/or allow distance calculations between stocks for a given sequence of days to find stocks or groups of stocks that tend to move together-- the stated goal for this analysis.

As such, a derived column for the data was produced with the following criteria:

1. IF a row of data has the same ticker symbol as the row before it
    a. To ensure comparisons only occur between the same stock
2. AND the row of data's value for close is greater than the row before it

. To enforce the rule that only days where the value has been increased DoD shall get a "1"

3. THEN the row shall get a 1 in a new column called "higherthanyesterday"

In order to accomplish this, the group built a python script utilizing Pandas and NumPy to do the associated csv extractions, calculations, and repackaging. The source code for the script was as follows:

```
#Group 2
#CSPB 4502 - Data Mining
#Final Project

#Import relevant libaries
import csv
import pandas as pd
import numpy as np

#Import reference csv
originaldata = pd.read_csv("fh_5yrs.csv")

#Extract symbol and close columns for analysis
symbols = originaldata["symbol"].to_numpy()
close = originaldata["close"].to_numpy()

#Create an list the length of the data and prefill with
increased = np.zeros(len(symbols))

#For each row, compare the previous row. If the symbol i
#has increased, change the 0 to a 1.
for i in range(1,len(symbols)):
    if symbols[i] == symbols[i-1] and close[i-1] < close
        increased[i] = 1

#Append to dataframe
originaldata["higherthanyesterday"] = increased

#Export as a csv
originaldata.to_csv("fh_5yrs_modified.csv")
```

Figure 1. Source code for "higherthanyesterday" derived attribute.

Following this, the resulting csv was opened in Excel to confirm the addition of the derived attribute column. As well, this additional column underwent extensive spot checking of the values to confirm script performance. Following these checks, it was determined that the script performed correctly and thus the augmented dataset is now ready for further analysis as the project progresses.

See below for a reference screenshot of the augmented csv, showing the additional column and correct values:

	date	volume	open	high	low	close	adjclose	syn
0	7/2/2020	257500	17.64	17.74	17.62	17.71	17.71	AA
1	7/1/2020	468100	17.73	17.73	17.54	17.68	17.68	AA
2	6/30/2020	319100	17.65	17.8	17.61	17.78	17.78	AA
3	6/29/2020	405500	17.67	17.69	17.63	17.68	17.68	AA
4	6/26/2020	335100	17.49	17.67	17.42	17.67	17.67	AA
5	6/25/2020	246800	17.6	17.6	17.52	17.59	17.59	AA
6	6/24/2020	329200	17.61	17.71	17.56	17.61	17.61	AA
7	6/23/2020	351800	17.55	17.66	17.55	17.66	17.66	AA
8	6/22/2020	308300	17.5	17.57	17.44	17.5	17.5	AA
9	6/19/2020	153800	17.27	17.4	17.26	17.4	17.4	AA
10	6/18/2020	102200	17.27	17.27	17.14	17.23	17.23	AA
11	6/17/2020	153900	17.19	17.25	17.16	17.25	17.25	AA

Figure 2. Snip of augmented csv opened in Excel.

A link to our updated data is as follows:

https://drive.google.com/drive/folders/17RyZ9pbuVMTZXn2SvbAhO8fvzA-08Kbv?usp=sharing

2. Number of Clusters

An important component of K-means clustering is to decide how many clusters are needed to accurately partition the data without overexplaining. Our goal was to maximize the bias-variance tradeoff using the elbow method. We measured the distortion for a several number of clusters and looked for the turning point of the graph. Distortion was measured as the sum of square distances from each point to its corresponding cluster center. Naturally, the distortion decreased with each additional cluster but with diminishing returns. The turning point, or elbow of

the graph, indicates the ideal number of clusters where any additional clusters do not explain enough of the variation in data to warrant additional complexity. With over 6000 stocks, there was a large possibility of optimal clusters. To narrow it down, we first looked at up to almost 200 clusters to get a general idea of the optimal k value.

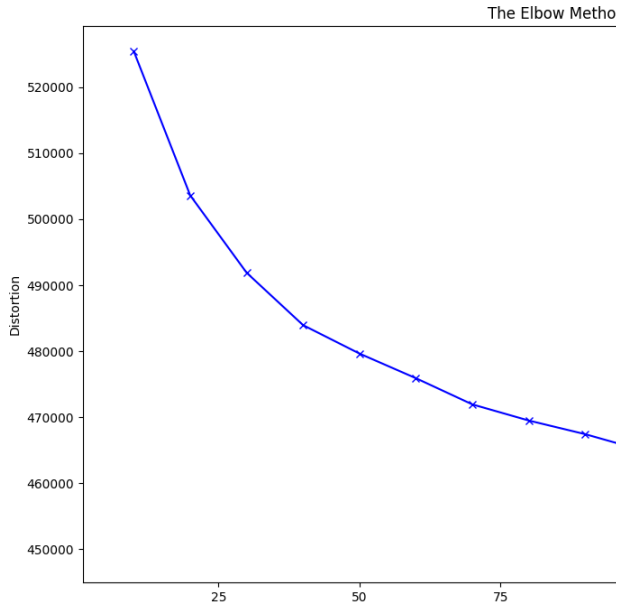

Figure 3. Elbow Method Results Measuring Distortion as a Function of K

The elbow point appeared to be between 25 and 50 clusters with a distortion between 480,000 and 490,000. We then took a closer look at the distortion number of k between 1-50.

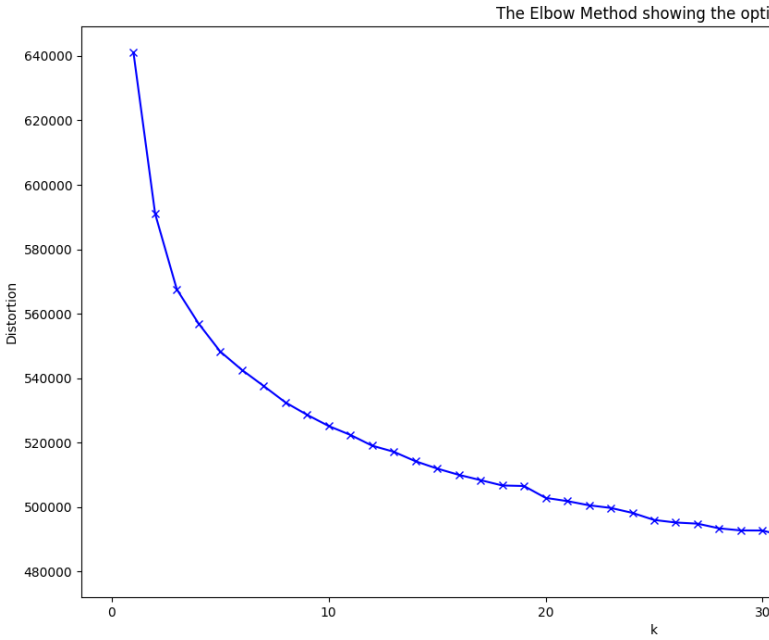

Figure 4. Elbow Method Results Measuring Distortion as a Function of K Ranging from 1-50 Clusters

This graph did not show a clear number of clusters as the elbow point. Figure 3 indicated the elbow point as somewhere around midway between 25 and 50. Since this method is heuristic and arbitrary, we chose a k essentially at the midpoint of 38. Figure 4 does not show clear evidence that there is a better number of clusters. In fact, 39 clusters even has a higher distortion score. Therefore, we decided to use 38 clusters as our optimal k in k-means clustering.

3. Resulting Cluster Analysis

Stocks are commonly grouped into different sectors in finance. These sectors are well-defined and intuitive. However, we chose to perform unsupervised learning instead to

discover subtle connections that might not be so intuitive. Our goal was to strike a balance. On one hand, completely random clusters would indicate a poor fitting model. At the other end of the spectrum, clusters that aligned perfectly with established sectors would not provide any new and interesting information. Ideally, we would get a result somewhere between these extremes.

As a sanity check, there appears to be intuitive clustering that indicates a model that fits. For example, cluster 13 includes oil and gas companies including BP, Chevron, and Exxon. Cluster 28 includes cinema entertainment companies like AMC, Dish, and IMAX. Cluster 35 includes healthcare companies like Pfizer, Johnson & Johnson, Merck and Cigna. With 38 clusters, the probability of these companies randomly falling into these respective clusters is negligible. The model appears to have intuitive sense.

As previously stated, clustering that ends up falling in line with established sectors is not only boring, but ineffective. We looked for interesting connections in the clusters. It is important to note that the clustering model only measures the connections between stock movement. The reasons behind these connections remain esoteric and we are not financial experts. Even financial experts often lament that making sense of the stock market is notoriously difficult. As data miners, we must abide by the cardinal rule that correlation does not imply causation. For example, Mcdonald's lies in cluster 35 while Wendy's lies in cluster 37. While both staples of the fast food industry, these two companies were not grouped together. It is possible that these two companies are in separate clusters because they are directly in competition. More customers at Mcdonald's means less people getting their hamburgers from Wendy's. Wendy's lies in the same cluster as Heinz Ketchup. Could they be linked because Wendy's uses more ketchup on their hamburgers than Mcdonald's and, therefore, is more closely linked to Heinz? Mcdonald's is in the same cluster as Lockheed Martin and Northrup Grumman. Do aerospace defense contractors prefer to take their lunch breaks at Mcdonald's over Wendy's? While these reasons may seem frivolous, there is no correct explanation because clustering does not explain how there are connections. It only explains that there are connections.

Our clustering did provide interesting connections beyond typical sector partitions. One example is the automobile industry. Ford and GM are located in cluster 32. Meanwhile, Toyota is located in cluster 16 while Honda is located in cluster 22. These results suggest that automobile stocks are not necessarily linked. At least, there is a stronger correlation than simply being car manufacturers. Ford and GM are American companies while Toyota and Honda are not. In addition, Ford and GM's cluster also includes American

steel production and mining companies such as United States Steel, Nucor, and Cleveland Cliffs. It appears that the success of Ford and GM are more closely tied to the success of the US steel industry than the automobile industry at large. After all, steel is the most commonly used metal in car production. This example shows that clustering stocks based on their industry sector is not necessarily the most accurate way to group stocks.

Another interesting result is cluster 28. It appears to be a cluster including several entertainment stocks such as AMC and Dish. It also appears to include several specialty retailers such as Gamestop, Build-a-Bear, Bed Bath and Beyond, and Tractor Supply Company. First of all, this cluster is a good example of the nature of our clustering. Since we categorized any increase in stock as the same, a drastic jump and a slight bump are considered the same in our analysis. Therefore, the drastic jumps in Gamestop's stock are irrelevant in our model. The only factor that matters is the direction of its daily movement and not the magnitude. Second of all, this cluster seems to include companies that are not essential for everyday life. The world does not need teddy bears, movies, or bath bombs to run. These are all luxury purchases. Therefore, the performance of this cluster could indicate the health of the economy. If consumers are spending money on movies and specialty retailers, they probably are making surplus money that exceeds their bills and necessities. In addition, this cluster could help classify companies as specialty retailers. For example, solar power companies like Sunpower and Canadian Solar are included in this cluster. This may indicate that solar energy is still considered a 'luxury' energy source. It could indicate that solar energy has not reached the same ubiquity as carbon energy sources and that considerable more work must be done to curb energy emissions.

In conclusion, we can posit our own explanations for the clusters but our clusters have the limitation of not providing the actual connections. We only know that these stocks move together and not why they do. Ultimately, we know the correlations but not the causations. We can make educated guesses but we do not have evidence for the reasons.This, however, is a feature rather than a bug. The stock market is notoriously difficult to predict and understand. With over 6,000 stocks in our analysis, we only recognized a few of them without further research. Investing in stocks typically requires considerable domain knowledge and financial expertise. Our clustering model simplifies the nebulous stock market. It does not require the knowledge of why stocks would be similar. In this way, our clustering model benefits from simplicity and makes the stock market more attainable for the amateur.

Applications

Our clustering model can be vital to investing. As mentioned above, it makes the stock market more accessible. Considerable domain knowledge is not needed for k-means clustering. This model only requires access to daily stock movement which is readily available. This could render expensive financial advisors and other experts as unnecessary.

A diversified portfolio is one where risk is hedged and minimized. A large loss in one portion of the portfolio is assuaged by a smaller loss or a gain in another portion of the portfolio. Since each cluster theoretically moves together, diversification can be as simple as selecting stocks in a variety of clusters. This method should be more effective than simply selecting stocks in different sectors. As shown in the Key Results section, Ford and GM were in separate clusters than Toyota and Honda. An investor could partially diversify their portfolio by investing in Ford, Toyota, and Honda. Likewise, one could invest in Ford and Nucor thinking they diversified by investing in a company in the automobile sector and a company in the materials sector. However, our clustering analysis showed that Ford and Nucor typically move in tandem. Our clustering model can allow for easy and effective diversification.

Another application is finding subtle links between different companies. This could be used to anticipate the trajectory of certain companies. A busy Build-a-Bear workshop could indicate that Gamestop will also experience success as they are both part of cluster 28. Likewise, increases in green energy could be correlated with the success of the solar power companies which lies in the same cluster 28 with the cinema entertainment industries. Does this mean blockbuster movies will decrease carbon emissions? It probably does not. However, people going to movies could indicate free time and spending money, which is indicative of a healthy economy. This, in turn, could show a willingness and ability to invest in high upfront costs like installing solar panels. Ultimately, the explanation behind the connections will need to be scrutinized and parsed together but our clustering does lay the groundwork for which companies are connected.

Further work can be done testing our clustering model. We could create a diversified portfolio based on the clustering and measure its gains compared to a diversified portfolio based on sectors, an unmanaged index fund, a financial advisor’s portfolio, random stocks, etc. We could even compare performances within our own model using a different number of clusters.

In conclusion, our original intention was to find unexpected and interesting connections between stocks. However, our clustering model could have several practical applications in investing. Further analysis and model comparison must be performed to ensure that our clustering model would be more

effective than other typical methods of creating a portfolio. Our K-means clustering provides considerable promise and a multitude of interesting connections to explore. This model can provide valuable insights into investing, economics, and the complicated connections between seemingly disparate companies.